

Escansión automática de poesía española sin silabación

Automatic Scansion of Spanish Poetry without Syllabification

Guillermo Marco Remón, Julio Gonzalo

nlp.uned.es, Research Group in NLP & IR, UNED, Madrid, Spain
{gmarco, julio}@lsi.uned.es

Resumen: En los últimos años, han surgido diversas herramientas de análisis métrico automático de poesía española. Estos sistemas se basan en complejos métodos de silabación y asignación de acentos, los cuales se apoyan en librerías de etiquetado gramatical, cuyo coste computacional es elevado. Este coste incrementa con el cálculo de ambigüedades métricas. El presente artículo parte de la hipótesis de que es posible llevar a cabo un análisis métrico informativo y preciso sin utilizar estos métodos. Se propone un algoritmo que realiza escansiones (número de sílabas, patrón métrico y tipo de verso) sin silabación. El algoritmo resuelve ambigüedades métricas y tiene en cuenta la compensación hemistiquial. Nuestros resultados indican que obtiene una mejora relativa sobre el estado del arte de un 2 % en la clasificación de patrones métricos en poesía de medida fija y un 25 % en poesía polimétrica. También se ejecuta 21 y 25 veces más rápido, respectivamente. Por último, se ofrece una aplicación de escritorio como herramienta para los investigadores de la poesía española.

Palabras clave: patrones métricos, escansión automática.

Abstract: In recent years, several systems of automated metric analysis of Spanish poetry have emerged. These systems rely on complex methods of syllabification and stress assignment, which use PoS-tagging libraries, whose computational cost is high. This cost increases with the calculation of metric ambiguities. However, it is possible to carry out an informative and accurate metric analysis without using these costly methods. We propose an algorithm that performs accurate scansion (number of syllables, stress pattern and type of verse) without syllabification. It addresses metric ambiguities and takes into account the hemistichs compensation. Our algorithm outperforms the current state of the art by 2 % in fixed-metre poetry, and 25 % in mixed-metre poetry. It also runs 21 and 25 times faster, respectively. Finally, a desktop application is offered as a tool for researchers of Spanish poetry.

Keywords: metrical patterns, automated scansion.

1 Introducción

En los últimos años, han surgido diversas herramientas de análisis métrico de poesía española. Dado un verso, extraen automáticamente el número de sílabas y su ritmo. Por ejemplo, a partir del verso:

- (1) *Amigos, el amor me perjudica*
A-mi-gos-el-a-mor-me-per-ju-di-
ca
 11 — 2.6.10
 (Julio Martínez Mesanza)

se determina que es un verso de once sílabas y su ritmo (también llamado esquema acentual o patrón métrico) es 2.6.10.

Para resolver este problema, los sistemas publicados hasta el momento se basan en complejos métodos de silabación y asignación de acentos, los cuales se apoyan en librerías de etiquetado gramatical, cuyo coste computacional es elevado. Como la escansión de un verso puede ser ambigua –los versos del Ejemplo 2 se pueden escandir a la manera (a) o a la manera (b)–, este coste incrementa con el

cálculo de las ambigüedades métricas.

(2) *dentro de su fluir los manantiales*

(a) *den-tro-de-su-fluir-los-ma-nan-tia-les*

10 — 1.5.9

(b) *den-tro-de-su-flu-ir-los-ma-nan-tia-les*

11 — 1,6,10

(Claudio Rodríguez)

Los versos por encima de las once sílabas se dividen en semiversos o hemistiquios para su escansión. Esta separación introduce fenómenos métricos que afectan al cómputo silábico, y por ello se han de considerar para la correcta determinación del ritmo. Los métodos actuales no tienen en cuenta estos fenómenos.

Este trabajo parte de la hipótesis de que es posible llevar a cabo un análisis métrico informativo y preciso sin utilizar métodos de silabación ni librerías de etiquetado gramatical. En la Sección 2 comenzaremos definiendo el problema de la escansión e introduciendo el marco teórico necesario para abordar su solución, y en la Sección siguiente resumiremos el estado del arte.

A partir del conocimiento métrico expuesto, en la Sección 4 se desarrolla el sistema análisis métrico automático que simplifica el problema de la medida del verso gracias a que prescinde de realizar silabación. El algoritmo tiene en cuenta la compensación hemistiquial y resuelve las ambigüedades métricas derivadas de sinalefas, dialefas, sinéresis y diéresis, sin perder precisión ni información sobre la decisión del analizador. La salida de este método ofrece el número de sílabas, el esquema métrico, el esquema métrico sin acentos extrarrítmicos, su nombre, su grado de coincidencia con el esquema métrico sin acentos extrarrítmicos, así como la forma del verso que ha llevado al sistema a clasificarlo en ese tipo: marca sinalefas, dialefas, sinéresis y diéresis.

En la Sección 5 se evalúa el sistema sobre conjuntos de datos de poesía de metro fijo y mixto. En nuestra experimentación, el algoritmo obtiene una mejora sobre el actual estado del arte de un 2 % en la clasificación de patrones métricos sobre poesía de medida fija y de un 25 % sobre poesía polimétrica. También se ejecuta 21 y 25 veces más rápido, respectivamente. Al final de la sección se

analizan los errores cometidos por el algoritmo.

Por último, en las Secciones 6 y 7 se describe *Jumper* una aplicación de escritorio que puede servir como herramienta para los investigadores de la poesía española, y se extraen conclusiones.

2 Definición del problema

Un verso es una serie de sílabas acentuadas y no acentuadas delimitada por pausas métricas (Caparrós, 2001). La distribución de los acentos sobre esta serie determina el patrón métrico. Así, la escansión consiste en clasificar los versos por su medida (número de sílabas) y su patrón métrico.

La sílaba es la unidad estructural de la palabra. En español, el núcleo de una sílaba es siempre vocálico (RAE, 2009, 8.1a). No obstante, la vocal también desempeña funciones de margen silábico en los diptongos; se forman combinando las vocales abiertas con cerradas, cerradas con abiertas, y las cerradas entre sí. Son un conjunto de 14 realizaciones: *ai, au, ei, eu, oi, ou, ui, iu, ia, ua, ie, ue, io, uo*.

En general, las palabras en español tienen sólo un acento, con algunas excepciones como los adverbios acabados en *-mente*. En función de dónde recae su acento, se clasifican y acentúan gráficamente de la siguiente manera:

- Palabras oxítonas o agudas: cuando la sílaba acentuada es la última de la palabra. Aunque existen algunas normas adicionales (RAE, 2010, 3.4.1.2.1), principalmente se acentúan con un signo gráfico cuando terminan en los grafemas de *n, s* o *vocal*.
- Paroxítonas o llanas: cuando la sílaba acentuada es la penúltima de la palabra. Se acentúan con un signo gráfico cuando terminan en un grafema distinto de *n, s* o *vocal*.
- Proparoxítonas o esdrújulas: cuando la sílaba acentuada es la antepenúltima de la palabra. Se acentúan con un signo gráfico siempre.

De manera homónima, un verso, por la posición de su última sílaba acentuada, recibe el nombre de oxítono, paroxítono o proparoxítono. De cada uno de ellos, resultan unos

fenómenos métricos constantes que afectan al número de sílabas (Quilis, 1984):

- Cuando un verso es oxítono, se cuenta una sílaba más.
- Cuando un verso es paroxítono, se cuenta las sílabas reales existentes.
- Cuando un verso es proparoxítono, se cuenta una sílaba menos.

Además de estas denominaciones, la tradición poética ha clasificado los versos en función de su patrón métrico. Un verso cuenta con un número finito de acentos. Un endecasílabo, por ejemplo, puede contar con cinco acentos rítmicos o menos; un octosílabo hasta cuatro. Por lo tanto, el número de combinaciones posibles también son finitas y asentadas en la tradición poética. Estos esquemas métricos tradicionales son los que el poeta tiene en mente cuando elige un metro. En el presente trabajo, se toma la clasificación de Pou (2020), por ser la más sistemática y exhaustiva. En la Tabla 1 se recoge la tipología del endecasílabo.

No obstante, en la realización de un verso, la distribución de los acentos puede ser variada y no coincidir exactamente con los ofrecidos en la clasificación. En el Ejemplo 3, se da el patrón métrico 1.6.(7).10, el cual es posible y frecuente en el endecasílabo, pero no se observa en la Tabla 1. Se advierte, sin embargo, que tiene los acentos característicos en 1.6.10 del ritmo enfático puro, siendo el acento en séptima extrarrítmico.

- (3) *Siempre la claridad viene del cielo*
Siem-pre-la-cla-ri-dad-vie-ne-del-cie-lo
 11 — 1.6.(7).10
 (Claudio Rodríguez)

Siguiendo este criterio, cada verso se puede clasificar en función de la proximidad a cada uno de los tipos: se observan los acentos característicos del tipo de verso y se tratan los no coincidentes como extrarrítmicos.

Cuanto mayor es el número de acentos, más complejo y rico es el verso. Sin embargo, cuando supera las once sílabas, se rompe en semiversos o hemistiquios. Por ejemplo, un verso de 14 sílabas, también llamado alejandrino, se suele dividir en hemistiquios de 7+7. Es decir, el verso en español tiene el límite rítmico del endecasílabo.

Endecasílabos		
— Heroicos	2.6.10	puro
	2.4.6.8.10	pleno
	2.4.6.10	corto
	2.6.8.10	largo
	2.4.10	difuso
— Melódicos	3.6.10	puro
	1.3.6.8.10	pleno
	3.6.8.10	largo
	1.3.6.10	corto
— Sáficos	4.8.10	puro
	1.4.8.10	puro pleno
	1.4.6.8.10	pleno
	4.6.10	corto
	1.4.6.10	corto pleno
	4.6.8.10	largo
	2.4.8.10	largo pleno
— Dactílicos	4.10	difuso
	1.4.10	difuso pleno
	4.7.10	puro
	1.4.7.10	pleno
— Enfáticos	2.4.7.10	corto
	1.6.10	puro
	1.6.8.10	pleno
— Vacíos	6.10	puro
	6.8.10	pleno
	1.4.10	pleno
	2.4.10	heroico

Tabla 1: Clasificación del verso de once sílabas (Pou, 2020, p. 118).

Precisamente, por esa subdivisión en hemistiquios, se puede realizar una escansión separada de cada uno de los semiversos. Así, en el Ejemplo 4 se contempla que el primer hemistiquio es proparoxítono, y resta una sílaba al cómputo. El segundo es oxítono y suma una. El resultado final es un verso alejandrino.

- (4) *Oh, qué frescor, qué música / de chopos de estación*
 $14 = 8 - 1 / 6 + 1$
 (Juan Ramón Jiménez)

La silabación es la segmentación de una palabra en sus sílabas constituyentes. Esta tarea viene determinada por la estructura morfológica del español (Española, 2009, 1.3.4a).

Es importante reparar en que la tarea de

silabación es distinta de la tarea de contar sílabas. Por ejemplo, dada la palabra *bolígrafo*, la silabación resultaría *bo-lí-gra-fo*, donde se ha hecho un esfuerzo morfológico por establecer la relación entre grafemas de la palabra. No obstante, el cómputo sílabico no informa de la relación entre los grafemas, sino que es una cuantificación: 4.

A parte de estas consideraciones generales, los versos son sometidos por los poetas a varios recursos métricos que afectan al cómputo sílabico e introducen ambigüedad en la determinación del patrón rítmico. Los más habituales son la sinalefa, dialefa, sinéresis y diéresis.

Sinalefa es la pronunciación en una sola sílaba de dos o más vocales que se encuentran en palabras distintas (Ejemplo 5).

- (5) *Creía que te había dicho adiós*
Cre-í-a-que-te-ha-bí-a-di-choa-diós
 11 — 2.6.8.10
 (Amalia Bautista)

En los versos de arte mayor superiores a once sílabas, la pausa hemistiquial rompe la sinalefa. En el Ejemplo 6 se rompe la sinalefa entre *mente* y *entre*.

- (6) *Escucho solamente entre las voces*
una
Es-cu-cho-so-la-men-te-entre-las-
vo-ces-u-na
 7 + 7 = 14
 (Antonio Machado)

La dialefa es la ruptura en dos sílabas de dos vocales de palabras distintas, es decir, lo contrario de la sinalefa.

La sinéresis consiste en la pronunciación en una sola sílaba de dos o más vocales que normalmente forman dos sílabas. “Hermosas ninfas que en el río metidas” (Garcilaso): “Hermosas ninfas que en el **rio** metidas” (11) (Pou, 2020, Capítulo III)

Por último, la diéresis consiste en la pronunciación de dos o más vocales que normalmente se realizan como un diptongo, convirtiéndolo en un hiato. Aunque suele venir explicitada por el poeta con el signo “. No siempre es así. Sirva de muestra el verso del Ejemplo 7, donde se encuentra una diéresis no indicada.

- (7) *Todas las tardes se muere un niño*

to-das-las-tar-des-se-mü-e-reun-
ni-ño

11 — 1.4.8.(9).10

(Federico García Lorca)

Además, la resolución de ambigüedades no es exacta. En el verso del Ejemplo 8, existen dos posibles escansiones, ambas válidas. En cualquier caso, de acuerdo con lo dicho anteriormente, se podría argumentar que el esquema 2.4.(7).8.10 con acento extrarrítmico en 7^a es natural del endecasílabo sáfico largo pleno y es, con un grado alto de certidumbre, el que el autor trataba de conseguir; por el contrario, el patrón 3.7.8.10 es artificial y su sonoridad está cuestionablemente alejada del endecasílabo. Hacia este tipo de decisiones se orienta el sistema presentado.

- (8) *Juez Elisio, que de un verde probo*
 (a) *Ju-ez-E-li-sio-que-de-un-ver-de-*
pro-bo
 11 — 2.4.(7).8.10
 (b) *Juez-E-li-sio-que-de-un-ver-de-*
pro-bo
 11 — 3.7.8.10
 (Lope de Vega)

3 Trabajos relacionados

La métrica, dentro de la filología, es un tema fundamental en el análisis literario de un poema. Entre los trabajos clásicos en el campo se encuentran los manuales de Bello (1859), Tomás (1956) y Quilis (1984). Más recientemente se publicó el diccionario y manual de Domínguez Caparrós (Caparrós, 1993; Caparrós, 2001). El presente trabajo se apoya en las últimas investigaciones de métrica española expuestas por Pou (2020)

En cuanto a los algoritmos de escansión automática, en los últimos años se encuentra el trabajo de Navarro-Colorado (2017). Se trata de un sistema basado en reglas que emplea el analizador morfológico Free-ling (Padró and Stanilovsky, 2012). Se centra en la resolución de ambigüedades; marca sinalefas y diéresis, pero no trata las sinéresis. Procesa las ambigüedades empleando una base de conocimiento de los distintos patrones métricos extraída en un análisis estadístico; aunque ofrece información adicional sobre el corpus, este análisis no es estrictamente necesario para realizar escansión, pues los patrones vienen ya dados por la propia preceptiva de

la métrica. Además, el sistema se evalúa sólo sobre versos de métrica fija.

Por otro lado, Agirrezabal, Alegria, and Hulden (2017) entrenan una red de neuronas bidireccional LSTM a nivel de carácter. El sistema predice patrones métricos a partir de una transformación enriquecida de la entrada, que incluye silabación. La salida no informa sobre la decisión del sistema, por lo que no se pueden conocer las ambigüedades que ha detectado. Se reportó una tasa de acierto de 0.91 sobre el corpus de metro fijo.

El sistema más reciente publicado es Rantanplan (de la Rosa et al., 2020). Se trata de un método basado en reglas que emplea un sistema de silabación del 99.99% de precisión. La señalización de acentos se apoya en esta silabación y en el etiquetado gramatical de la librería Spacy (Honnibal and Montani, 2017). Como indican los autores, el sistema está limitado por este modelo estadístico: comete errores ocasionales asignando los acentos y, desde el punto de vista del coste computacional, tan sólo su carga retrasa 18 segundos el análisis de acuerdo con las pruebas de los investigadores. No tienen en cuenta los efectos de compensación y diálefa hemistiquiales, lo cual perjudica la tasa de acierto (*accuracy*) sobre el corpus de poemas polimétricos.

Todos los métodos anteriores utilizan de una manera u otra métodos de silabación. Tampoco tienen en cuenta los fenómenos métricos de los versos de más de once sílabas.

4 *Jumper: algoritmo de análisis métrico sin silabación*

En esta Sección se presenta *Jumper*, el método de análisis métrico automático desarrollado. Se divide a su vez en tres subsecciones: análisis de palabra, verso y poema.

4.1 Módulo 1: Análisis de palabra: cómputo de sílabas y acentos

Como se ha dicho en la Introducción, en español, el núcleo de una sílaba es siempre vocálico. La unidad de la sílaba viene dada por la vocal. Por lo tanto, para contar sílabas de manera eficiente, basta con contar las vocales de una palabra, teniendo en cuenta los diptongos. De hecho, no es necesario localizar los hiatos explícitamente: los diptongos en español son un conjunto finito de 14 realizaciones. Por lo tanto, dada la aparición

contigua de dos vocales, basta comprobar si es un diptongo para no contar esa siguiente vocal como sílaba.

Además, este mismo procedimiento permite localizar el acento. Se asume que la palabra sigue las reglas de acentuación gráfica del español (RAE, 2010, 3.4). Así, mientras se cuentan las vocales, se guarda la posición en la que se ha encontrado una vocal acentuada gráficamente. Si tras el cómputo no se ha encontrado el acento, se aplican las reglas de acentuación básicas del español para hallarlo.

En cuanto se conoce si la palabra es aguda, llana o esdrújula, se puede calcular fácilmente la compensación cuando se encuentra al final del verso.

Gracias a este método, al tener en cuenta la vocal acentuada explícitamente, se puede también localizar el acento cuando el poeta emplea el recurso de la sístole.¹

Cabe decir que, de momento, se está operando a nivel de palabra. En cuanto se traten de resolver las ambigüedades métricas de un verso producidas por las sinéresis sí será necesario dar un tratamiento específico a los hiatos.

4.2 Módulo 2: Análisis de verso

Una vez se ha diseñado la función que devuelve el número de sílabas y el lugar de sus acentos individualmente, se quiere extender el análisis al verso entero. Este módulo es el núcleo del método, y es el que entraña más complejidad. Se divide, a su vez, de tres submódulos. El primero se encarga del cómputo de sílabas y acentos; el segundo, de la compensación hemistiquial para los versos mayores de once sílabas y, el tercero, de resolución de ambigüedades métricas.

4.2.1 Submódulo 1: Cómputo de sílabas y acentos de un verso

El primer submódulo funciona de la siguiente manera. Dado un verso, se convierte en una lista de palabras. Para cada palabra de la lista, se calculan, con la función expuesta en la Sección 4.1, el número de sílabas, el lugar del acento en esas sílabas y el factor de compensación. Este factor es -1 para las palabras esdrújulas, 0 para las llanas y 1 para las agudas. Se suma el número de sílabas de la palabra al número total de sílabas del

¹Sístole consiste en adelantar una o más sílabas el acento normal de una palabra; jilguero - jílguero. (Pou, 2020, Capítulo III)

verso, teniendo en cuenta, si la hubiera, la sinalefa con la palabra siguiente. Se calcula la posición del acento y se añade a la lista del patrón métrico. Si la palabra es un adverbio en *-mente* se tienen en cuenta sus dos acentos. Si la palabra es átona, no se añade a la lista de acentos.

En este primer submódulo se calcula la medida del verso teniendo en cuenta todas las sinalefas. Posteriormente, si hay alguna que no habría que tener en cuenta, se resuelve en el módulo de ambigüedades.

4.2.2 Submódulo 2: Compensación hemistiquial

Al comienzo del análisis de un verso, no se conoce cuántas sílabas va a tener. Si tras el cómputo del primer submódulo resulta un verso de más de once sílabas, se hace una llamada recursiva sobre la función indicando que ahora se tenga en cuenta la compensación hemistiquial, activando así el segundo submódulo. Si durante este segundo cómputo se hallan hemistiquios, se tiene en cuenta el factor de compensación de cada semiverso y la ruptura de la sinalefa por la pausa hemistiquial, si la hubiera. Se considera que una palabra se encuentra en la frontera de un hemistiquio cuando el número de sílabas del verso en proceso sumado al número de sílabas de la palabra y su factor de compensación resultan la medida esperada del hemistiquio (por ejemplo, 7 para el alejandrino).

4.2.3 Submódulo 3: Resolución de ambigüedades

Para calcular correctamente el patrón métrico de un verso, hay que tener en cuenta los recursos métricos empleados en él. Es necesario detectar dialefas, sinéresis y diéresis, sus posibles combinaciones, y obtener el mejor candidato.

Por lo tanto, el submódulo de resolución de ambigüedades tiene dos tareas. La primera consiste en detectar las posibles ambigüedades del verso y generar candidatos con ellas. La segunda consiste en elegir, entre los posibles candidatos, el mejor.

La activación del submódulo de detección de ambigüedades se indica desde el módulo superior de análisis de poema. Una vez indicada, por eficiencia, la detección se hace al mismo tiempo que el cómputo del Submódulo 1. Cuando se detecta una sinalefa, se genera a la vez un candidato con dialefa. Cuando se detecta un diptongo, se genera un candidato

con diéresis. Cuando se detecta un hiato, se genera un candidato con sinéresis. Una vez se han almacenado todos los candidatos con las posibles ambigüedades, se combinan: por ejemplo, un verso puede presentar, al mismo tiempo, una dialefa y una sinéresis. Todos los candidatos se etiquetan con su recurso métrico para así conservar la decisión del submódulo de resolución de ambigüedades.

En la segunda tarea, se elige el mejor candidato. Como se ha expuesto en la introducción, los patrones métricos de los versos son finitos y sus posiciones se aproximan a los tipos de versos asentados en la tradición. Por lo tanto, se compara la proximidad del vector del patrón rítmico extraído con cada uno de los tipos asentados en la preceptiva literaria; en función de esa proximidad, se elige el mejor candidato.

Aunque probar todas las posibles ambigüedades podría parecer costoso computacionalmente, es rápido por dos razones. La primera es que los versos ambiguos son sólo un pequeño subconjunto del corpus. La segunda es que, gracias al sencillo algoritmo de cómputo de sílabas y acentos, el coste de volver a computar un verso es muy bajo.

4.3 Módulo 3: Análisis de poema

Dado un poema, se convierte en una lista de versos. Se realiza la escansión de cada uno de ellos por medio del módulo de análisis a nivel de verso, con el sistema de resolución de ambigüedades desactivado.

Una vez calculadas todas las medidas y patrones rítmicos, se observa la frecuencia de los números de sílabas. Pueden darse dos casos. El primero, cuando sólo hay una medida frecuente: en este caso, se concluye que es un corpus de metro fijo. Si la lista de medidas frecuentes es mayor de uno (por ejemplo, en la poesía contemporánea, son habituales los poemas de versos polimétricos de 7, 11, y 14 sílabas), se concluye que es un corpus de poemas polimétricos. Una vez se tiene la lista de medidas frecuentes, se vuelven a computar los versos que no coincidan con ellas, ahora con el módulo de resolución de ambigüedades activado. La única diferencia entre poesía de metro fijo y mixto es que, para esta última, la lista de medidas frecuentes se actualiza en un contexto de un número n de versos. El tamaño del contexto es el único hiperparámetro del sistema. Por defecto, se establece en 14, ya que el número de versos del poema por

antonomasia, el soneto.

Una vez resueltas las ambigüedades, se almacena en una tabla cuyas columnas son: verso, verso etiquetado, número de sílabas, patrón métrico en forma de vector de enteros, patrón métrico en forma de vector de enteros del tipo de verso sin acentos extrarrítmicos, nombre del verso y, por último, ratio de coincidencia entre el patrón calculado y el patrón sin acentos extrarrítmicos. La salida del sistema se muestra en la Figura 1.

5 Evaluación

En los últimos años se ha estandarizado el uso del corpus anotado de sonetos (estrofa de 14 versos de once sílabas) del Siglo de Oro (Navarro-Colorado, Lafoz, and Sánchez, 2016) para la evaluación de los sistemas de escansión sobre poemas de medida fija. Originalmente, consistía en un conjunto de 1400 versos etiquetados manualmente con su patrón métrico; posteriormente se ampliaron a 10268. Está disponible en la herramienta de descarga de corpus Averell del proyecto POSTDATA.²

Asimismo, para métrica de medida mixta, de la Rosa et al. (2020) introdujeron recientemente un corpus elaborado a partir de la antología de Antonio Carvajal (1983). Se compone de 4378 versos, tanto de arte mayor como de arte menor, etiquetados manualmente. Por razones de derechos de autor, no es de dominio público; no obstante, sus desarrolladores nos permitieron su uso para esta experimentación.

Los sistemas se evalúan en función de su acierto (*accuracy*) establecido de forma binaria para cada verso: el acierto es 1 si se han identificado correctamente todos los acentos anotados del verso, y 0 en caso contrario. El rendimiento del sistema se mide como la tasa de acierto sobre el total de versos del corpus.

Todos los experimentos de nuestro sistema se han realizado sobre un ordenador equipado con la misma configuración que se empleó para las evaluaciones de los distintos métodos comparados en Rantanplan: procesador Intel® Core™ i7-8550U CPU @ 1.80GHz y 16GiB de memoria RAM DDR4.

El acierto (*accuracy*) se reporta en el intervalo [0,1] con dos cifras significativas. El tiempo se indica en segundos.

Al sistema se le ha dado el nombre de *Jumper*.

5.1 Evaluación sobre poemas de medida fija

La Tabla 2 recoge la tasa de acierto (*accuracy*) y el tiempo de cada uno de los métodos sobre el corpus de 10268 versos de medida fija elaborado por Navarro-Colorado, Lafoz, and Sánchez (2016). Nuestro sistema, *Jumper*, obtiene una tasa de acierto de 0,95, lo que supone una mejora del 2,2 % respecto al actual estado del arte establecido por Rantanplan. Además, se ejecuta 21 veces más rápido que éste.

Estos resultados indican que prescindir de silabación supone una mejora en el enfoque de la solución al problema. Gracias a que no se requiere la compilación de numerosas expresiones regulares para la separación de sílabas ni los pesados modelos de PoS-Tagging para la asignación de acentos, el tiempo es menor. La velocidad conseguida permite realizar una desambiguación exhaustiva en poco tiempo, lo que influye también en la mejora de la tasa de acierto.

Método	Accuracy	Tiempo
Navarro-Colorado	0,91	16787s
Rantanplan	0,93	53s
Jumper (nuestro)	0,95	2,5s

Tabla 2: *Accuracy* y tiempo sobre el corpus de 10268 versos de métrica fija de Navarro-Colorado. Los resultados de Navarro-Colorado y Rantanplan se han extraído de de la Rosa et al. (2020). Los de *Jumper* se han obtenido bajo la misma configuración de hardware.

Es interesante examinar el resultado de la evaluación sobre el subconjunto inicial de 1400 versos establecido por sus autores, ya que es el único corpus para esta tarea sobre el que se ha reportado una *inter-annotator agreement* (IAA) o tasa de acuerdo entre anotadores. La IAA sobre este subconjunto es de 0,96 (Navarro-Colorado, Lafoz, and Sánchez, 2016). *Jumper* obtiene un 0,95. No obstante, se ha detectado en el análisis de errores de la Sección 5.3.1 un sesgo en la anotación de estos primeros 1400 versos: la interjección “oh”, frecuente en el Siglo de Oro, es considerada átona sólo en este subconjunto del

²<https://github.com/linhd-postdata/averell>

Análisis							
N	Verso	Etiquetado	Sílabas	Acentos	Sin extrarrítmicos	Tipo	Coincidencia
1	Decir pestes de él tiene, sin duda,	decir pestes de él tiene sin duda	11	[2, 3, 6, 7, 10]	[2, 6, 10]	Endecasílabo heroico puro	80
2	un sólido prestigio literario	un sólido prestigio literario	11	[1, 2, 6, 10]	[2, 6, 10]	Endecasílabo heroico puro	90
3	-tacharlo de asesino, por ejemplo,	-tacharlo de asesino, por ejemplo,	11	[2, 6, 10]	[2, 6, 10]	Endecasílabo heroico puro	100
4	o compararlo con	o compararlo con	7	[4, 6]	[4, 6]	Heptasílabo sáfico puro	100
5	uno de esos cícloes con nombre de corista	uno de esos cícloes con nombre de corista	14	[1, 3, 6, 9, 13]	-	Aleandrino	100
6	que pasan y que dejan en los teledí-arios	que pasan y que dejan en los teledí-arios	14	[2, 6, 13]	-	Aleandrino	100
7	un paisaje de grandes palmeras derrocadas	un paisaje de grandes palmeras derrocadas	14	[1, 3, 6, 9, 13]	-	Aleandrino	100
8	y urallitas errantes,	y urallitas errantes,	7	[3, 6]	[3, 6]	Heptasílabo melódico puro	100
9	o simplemente lamentarlo a base	o simplemente lamentarlo a base	11	[2, 4, 8, 10]	[2, 4, 8, 10]	Endecasílabo sáfico largo pleno	100
10	de tardes y de otoños en pálidos jardines,	de tardes y de otoños en pálidos jardines,	14	[2, 6, 9, 13]	-	Aleandrino	100
11	pero ahora, con la mano en el poema,	pero ora con la mano en el poema	11	[2, 6, 10]	[2, 6, 10]	Endecasílabo heroico puro	100
12	os lo confieso: he sido siempre yo	os lo confieso: he sido siempre yo	11	[4, 5, 6, 8, 10]	[4, 6, 8, 10]	Endecasílabo sáfico largo	90
13	el que salió ganando de todos nuestros tratos.	el que salió ganando de todos nuestros tratos.	14	[4, 6, 9, 13]	-	Aleandrino	100

Figura 1: Salida del sistema. Obsérvese la diéresis detectada en el décimo verso.

corpus. Así, únicamente añadiendo a la lista de palabras átonas esa interjección Jumper obtendría un 0,96. Por tanto, nuestro sistema alcanza el límite de precisión que se puede medir con este corpus, ya que iguala la tasa de acuerdo entre anotadores. Este conjunto también permite comparar el efecto del tiempo de arranque de los sistemas sobre conjuntos de datos de tamaño pequeño. Bajo la misma configuración, Jumper analiza los 1400 versos en 0,33 segundos frente a los 2356 del sistema de Navarro-Colorado y los 21 segundos de Rantanplan.

5.2 Evaluación sobre poemas de medida mixta

En la Tabla 3 se contempla la tasa de acierto (*accuracy*) sobre el corpus de métrica mixta. Jumper obtiene 0,82, lo que supone un 25 % de mejora relativa respecto al actual estado del arte establecido por Rantanplan. En cuanto al tiempo, se ejecuta 25 veces más rápido que éste.

Método	Accuracy	Tiempo
Navarro-Colorado	0,49	7484s
Rantanplan	0,65	27s
Jumper (nuestro)	0,82	1,1s

Tabla 3: *Accuracy* y tiempo de ejecución sobre el corpus de 4300 versos de métrica mixta de Antonio Carvajal. Los resultados de Navarro-Colorado y Rantanplan se han extraído de de la Rosa et al. (2020), donde ambos han sido ejecutados en el mismo entorno. Los de Jumper se han obtenido bajo la misma configuración de hardware.

Estos resultados son producto de los beneficios de prescindir de silabación y librerías de etiquetado gramatical, ya comentados en la Sección anterior. Pero, además, se añade

otro factor en la mejora sobre este corpus: la consideración de los fenómenos métricos particulares de los versos de más de once sílabas. Esta consideración permite clasificar sus patrones métricos con mayor acierto. En el corpus de medida mixta de Carvajal se encuentran mezclados tanto versos de 7, 9 u 11 como de 14. El Ejemplo 9, tomado del conjunto de datos, Jumper lo clasifica correctamente como verso de 14 sílabas con ritmo 1.3.6.8.10.13, ya que el primer hemistiquio es oxítono y suma una sílaba. Rantanplan, sin embargo, lo clasifica como verso tridecasílabo de ritmo 1.3.6.7.9.12, al no tener en cuenta la compensación. Lo mismo ocurre con el fenómeno de ruptura de sinalefa entre hemistiquios.

(9) *una lucha común, y un descanso común*

- (a) $6+1 / 6+1 = 14$ — 1.3.6.8.10.13
(Jumper)
- (b) 13 — 1.3.6.7.9.12 (Rantanplan)
(Antonio Carvajal)

5.3 Análisis de errores

En esta Sección se profundiza en las circunstancias en las que Jumper falla en su clasificación de patrones métricos. En el corpus de medida fija se clasifican incorrectamente 546 de 10268 patrones. Sobre el corpus de medida mixta se clasifican incorrectamente 800 de 4378.

5.3.1 Errores sobre medida fija

Se ha tomado una muestra de 100 versos de los 546 clasificados incorrectamente. Tras su análisis manual, se contempla que hay principalmente cuatro fuentes de error.

La primera fuente de error (35 equivocaciones), atribuible al sistema, es la elección incorrecta del mejor candidato entre las posibles realizaciones de un verso ambiguo. Cuando un verso tiene dos escansiones posibles

igual de correctas, ambas con el mismo ratio de coincidencia con el patrón asentado en la tradición, la resolución de este empate no siempre se hace correctamente. Dado el verso del Ejemplo 10, el sistema devuelve el ritmo 3.6.8.10, que es igual de correcto que el anotado, 2.6.8.10. Sin embargo, cuando el acento en la sexta sílaba recae en un diptongo “cruel”, la diéresis “crü-el” es más natural que la dialefa entre “si” y “a” elegida por Jumper.

(10) *si a Silvia la cruel pastora viere*

- (a) *si-a-Sil-via-la-cruel-pas-to-ra-vie-re*,
11 — 3.6.8.10 (Jumper)
- (b) *sia-Sil-via-la-crü-el-pas-to-ra-vie-re*
11 — 2.6.8.10 (Anotado)
(Hernando de Acuña)

La segunda fuente de error es la anotación manual incorrecta, que se detecta en 26 de los 100 versos analizados. En algunas ocasiones, los errores se deben a desambiguaciones imprecisas del anotador. El Ejemplo 11 se ha anotado con el ritmo 2.3.7.10, el cual es extraño al endecasílabo. La escansión 2.(3).6.10 del endecasílabo heroico puro es la correcta para ese verso.

(11) *dolor pide a Felipe de Lianño*

- (a) *do-lor-pi-de-a-Fe-li-pe-de-Li-a-ño*
11 — 2.3.6.10 (Jumper)
- (b) *do-lor-pi-de-a-Fe-li-pe-de-Lia-ño*
11 — 2.3.7.10 (Anotado)
(Lope de Vega)

La tercera fuente de error (21 equivocaciones), también relacionada con la anotación, se debe a la consideración de palabras átonas como tónicas y viceversa. Por ejemplo, en los primeros 1400 versos se suele considerar átona la interjección “oh”, mientras que, en los restantes, se considera tónica. Jumper la considera tónica. Solo debido a la arbitrariedad de la acentuación de esta interjección, tan frecuente en el Siglo de Oro, se detectan en el análisis manual 11 errores.

La cuarta fuente de error son 8 versos con errores ortográficos que provocan asignaciones de acentos incorrectas.

La atribución de los 10 errores restantes es ambigua.

5.3.2 Errores sobre medida mixta

Se ha revisado manualmente una muestra de 100 de los 800 versos clasificados incorrectamente por Jumper.

Se han hallado 21 errores de anotación. Es importante recalcar que el corpus de Carvajal ha sido recientemente introducido para esta tarea y se encuentra todavía en un proceso de refinamiento. De hecho, la herramienta puede contribuir a limpiar la anotación, puesto que si se revisan los errores de Jumper se encuentra que uno de cada cinco casos es atribuible a un error manual de anotación y no a un error del sistema.

Otra fuente de error es que en ocasiones los anotadores no consideran tónicos algunos determinantes indefinidos y demostrativos: es el caso de “un, una, unos, unas, este, esta...”. La mayor parte de las veces, estos determinantes se encuentran en acentos extrarrítmicos, de modo que se anulan por el acento rítmico contiguo. Sin embargo, es un acento que realmente se encuentra en el verso, y aunque posteriormente en un análisis abstracto se elimine, consideramos que se ha de tener en cuenta en la escansión. Como Jumper, de acuerdo con RAE (2009, 9.7b), sí los considera acentuados, falla en 46 de los 100 versos analizados por esta razón. Sin embargo, como hemos explicado, podría considerarse también un error de anotación. En nuestras pruebas comprobamos que si se añaden estos determinantes tónicos a la lista de palabra átonas, el acierto sobre este corpus baja del 0,82 al 0,76, de lo que se deduce que los anotadores en algunas ocasiones consideran estos determinantes átonos y en otras ocasiones tónicos.

En menor medida (6 casos), se han hallado errores derivados de no emplear etiquetado gramatical. Por ejemplo, cuando se procesa la palabra “mientras” nuestro sistema la considera átona, ya que su función más frecuente es conjuntiva (“Estudia *mientras* yo leo”) (Quilis, 1984, p. 25); sin embargo, cuando tiene función de adverbio, es tónica “estudia; *mientras*, yo leo”. Este problema no es una dificultad insalvable con el sistema implementado, ya que se pueden resolver este tipo de errores tratándolos como una ambigüedad más. En cualquier caso, la frecuencia tan baja de errores atribuibles a la falta de etiquetado gramatical confirma que la estrategia de Jumper es preferible.

Del mismo modo que en el corpus de medi-

da fija, el sistema falla 23 veces por no elegir correctamente el mejor candidato entre los versos ambiguos.

Los 4 casos restantes se deben a errores ortográficos que provocaban asignaciones de acentos incorrectas.

6 Interfaz gráfica para Jumper

Se ha desarrollado una aplicación de escritorio para ofrecer una interfaz al método expuesto. Mientras el usuario escribe el poema, se ejecuta el algoritmo y se imprime el análisis en tiempo real. Para cada verso, si no tiene acentos extrarrítmicos se imprime en verde. Si no coincide plenamente con el esquema acentual sin extrarrítmicos, se imprime en negro. Si no cumple la tendencia versal del poema, en rojo. La tendencia versal se calcula de forma automática si el usuario no la introduce explícitamente. Esta aplicación puede ser de utilidad para investigadores de la poesía española y traductores. Puede descargarse en el siguiente enlace: <https://github.com/grmarco/jumper>

7 Conclusiones

En este trabajo se ha presentado un algoritmo de escansión automática de poemas en español sin necesidad de silabación. A partir del conocimiento de la métrica española, se han establecido una serie de premisas que han simplificado el problema. De manera resumida, las premisas son:

- La tarea de silabación es distinta a la de contar sílabas.
- La vocal es núcleo de la sílaba en español; por lo tanto, es lo que identifica su unidad. Así, se ha podido desarrollar un sencillo algoritmo de cómputo de sílabas y acentos (Sección 4.1), que ha mejorado notablemente la eficiencia sin necesidad de uso de librerías externas.
- Ante un verso ambiguo, las realizaciones posibles del patrón métrico son finitas y asentadas en la tradición. Por lo tanto, se trata de resolver las ambigüedades teniendo en cuenta la aproximación a los patrones naturales del verso.
- La compensación y dialefa hemistiquial se han de tener en cuenta para realizar escansiones precisas.

Gracias a estas premisas, se ha desarrollado un sistema análisis métrico automático que simplifica el problema de la medida del verso, tiene en cuenta la compensación hemistiquial y resuelve las ambigüedades métricas derivadas sinalefas, dialefas, sinéresis y diéresis, sin perder precisión ni información sobre la decisión del analizador.

Nuestro algoritmo, Jumper, mejora el actual estado del arte en un 2 % para la clasificación de patrones métricos sobre poesía de medida fija, y en un 25 % sobre poesía de medida mixta. Además, todas las evaluaciones se han ejecutado entre 21 y 25 veces más rápido que el estado del arte.

También se ha llevado a cabo un análisis de los errores cometidos por el sistema, lo que permite vertebrar trabajos futuros para solucionarlos, y también depurar los corpus utilizados, ya que entre 1/5 y 1/4 de los supuestos errores del sistema son en realidad problemas de la anotación manual.

Finalmente, se ha desarrollado una interfaz gráfica para el algoritmo de análisis métrico en tiempo real, que puede ser de utilidad para investigadores de poesía española.

Agradecimientos

Esta investigación se ha desarrollado gracias al proyecto MISMI-BIAS (PGC2018-096212-B-C32), financiado por el Gobierno de España, Ministerio de Ciencia, Innovación y Universidades.

Para la evaluación del sistema expuesto han sido de ayuda la facilitación de los corpus empleados por parte de los autores de Rantanplan (de la Rosa et al., 2020).

References

- Agirrezabal, M., I. Alegria, and M. Hulden. 2017. A comparison of feature-based and neural scansion of poetry. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, Ranlp 2017*, pages 18–23.
- Bello, A. 1859. *Principios de la ortología i métrica de la lengua castellana*. La Opinión.
- Caparrós, J. D. 1993. *Métrica española*. Síntesis Madrid.
- Caparrós, J. D. 2001. *Diccionario de métrica española*. Alianza.
- Carvajal, A. 1983. *Extravagante jerarquía: 1968-1981*, volume 58. Hiperión.

- de la Rosa, J., Á. Pérez, L. Hernández, S. Ros, and E. González-Blanco. 2020. Rantanplan, fast and accurate syllabification and scansion of spanish poetry. *Procesamiento del Lenguaje Natural*, 65:83–90.
- Española, R. A. 2009. *Nueva gramática de la lengua española*, volume 2. Espasa Libros.
- Honnibal, M. and I. Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*, 7.
- Navarro-Colorado, B. 2017. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.
- Navarro-Colorado, B., M. R. Lafoz, and N. Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364.
- Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *International Conference on Language Resources and Evaluation*.
- Pou, P. J. 2020. *Métrica española*. Ediciones Cátedra.
- Quilis, A. 1984. *Métrica española*. Ariel Barcelona.
- RAE. 2009. *Nueva gramática de la lengua española. Fonética y fonología*, volume 3. Espasa Libros.
- RAE, R. A. E. 2010. *Ortografía de la lengua española*. Espasa.
- Tomás, T. N. 1956. *Métrica española*, volume 4. Las Americas Publishing Company.

A Apéndice: Código fuente y reproducibilidad

Se pueden reproducir los resultados del analizador en este repositorio: <https://github.com/grmarco/jumper-evaluation>